مجله تخصصی اپیدمیولوژی ایران؛ زمستان ۱۳۹۷، دوره ۱٤، شماره ٤: صفحات ۳۸۳-۳۷۵.

مقاله پژوهشی

مقایسه روش‌های جنگل تصادفی و رگرسیون لجستیک در پیش‌بینی مرگ‌ومیر مبتلایان به سرطان کولورکتال و عوامل مرتبط با آن

علی محمد کشت ورز حسام‌آبادی[۱]، ابراهیم حاجی‌زاده[۲]، محمدامین پورحسینقلی[۳]، احسان ناظم الحسینی مجرد[٤]

[۱] دانشجوی کارشناسی ارشد آمار زیستی، دانشکده علوم پزشکی، دانشگاه تربیت مدرس، تهران، ایران

[۲] استاد، گروه آمار زیستی، دانشکده علوم پزشکی، دانشگاه تربیت مدرس، تهران، ایران

[۳] دانشیار، مرکز تحقیقات گوارش و کبد، پژوهشکده بیماری‌های گوارش و کبد، دانشگاه علوم پزشکی شهید بهشتی، تهران، ایران

[۴] دکترای پزشکی مولکولی، مرکز تحقیقات گوارش و کبد، پژوهشکده بیماری‌های گوارش و کبد، دانشگاه علوم پزشکی شهید بهشتی، تهران، ایران

نویسنده رابط: ابراهیم حاجی زاده، نشانی: تهران، جلال آل احمد، پل نصر، دانشگاه تربیت مدرس، دانشکده علوم پزشکی، گروه آمار زیستی، تلفن: ۸۲۸۸۳۸۱۰.

پست الکترونیک: hajizadeh@modares.ac.ir

تاریخ دریافت: ۹۷/۰۳/۱۱؛ پذیرش: ۹۷/۰۷/۰۷

مقدمه و اهداف: هدف این مطالعه پیش بینی مرگ ومیر حاصل از سرطان کولورکتال در بیماران ایرانی و تعیین عوامل موثر بر آن با استفاده از روش جنگل تصادفی و رگرسیون لجستیک است.

روش کار: از اطلاعات ۳۰۴ بیمار مبتلا به سرطان کولورکتال ثبت مرکز تحقیقات گوارش و کبد دانشگاه علوم پزشکی شهید بهشتی طی سال‌های ۸۸ تا ۹۳ به صورت یک مطالعه گذشته نگر استفاده شد. تجزیه و تحلیل داده‌ها با استفاده از روش‌های جنگل تصادفی و رگرسیون لجستیک انجام شد. برای تحلیل داده‌ها از نرم افزار R نسخه ۳ .۴ .۳ استفاده شد.

یافته‌ها: ده متغیر مهمی که با مرگ ومیر سرطان کولورکتال ارتباط دارند، توسط روش جنگل تصادفی انتخاب شدند. چندین معیار مانند مساحت زیر منحنی مشخصه عملکرد (AUC) برای مقایسه روش جنگل تصادفی با رگرسیون لجستیک در نظر گرفته شد. با توجه به هر دو معیار، پنج متغیر اثر گذار رتبه بندی شده توسط جنگل تصادفی عبارتند از: مرحله سرطان، سن تشخیص بیماری، سن بیمار، مکانیسم فرار از سلول ایمنی و درجه تمایز یافتگی تومور. از نظر معیارهای مختلف، روش جنگل تصادفی عملکرد بهتری نسبت به رگرسیون لجستیک داشت (مساحت زیر منحنی ROC به ترتیب برای روش جنگل تصادفی و رگرسیون لجستیک برابر با ۹۸٪، ٪۸۰).

نتیجه‌گیری: متغیرهای سن تشخیص، مرحله سرطان، سن بیمار، مکانیسم فرار از سلول ایمنی و درجه تمایز یافتگی تومور به عنوان مهمترین عوامل موثر بر مرگ ومیر در سرطان کولورکتال به حساب می‌آیند که با تشخیص زودرس سرطان با برنامه های بیماریابی و غربالگری می‌توان بر طول عمر بیماران افزود.

واژگان کلیدی: سرطان کولورکتال، جنگل تصادفی، رگرسیون لجستیک، مرگ‌ومیر

مقدمه

سرطان یکی از عوامل اصلی مرگ‌ومیر در جهان است که به‌عنوان یک مسئله مهم و قابل‌توجه در حوزه بهداشت عمومی شناخته می‌شود (۱). این بیماری دومین علت مرگ‌ومیر در کشورهای توسعه‌یافته و سومین عامل مرگ‌ومیر در کشورهای کمتر توسعه‌یافته است (۲).

سرطان در ایران پس از بیماری‌های قلبی-عروقی و سوانح، سومین عامل مرگ‌ومیر به شمار می‌آید. بر طبق گزارش‌های سالانه مرکز ثبت سرطان در ایران، سرطان کولورکتال پس از سرطان‌های معده، مثانه و پروستات در مردان، چهارمین سرطان است و در زنان بعد از سرطان سینه، دومین سرطان شایع محسوب می‌شود (۴, ۳).

در ایران هرسال سرطان کولورکتال، بروزی حدود ۶ تا ۷/۹ در ۱۰۰ هزار نفر جمعیت دارد (۵) و با میزان مرگی حدود ۱۱۹۸ در ۱۰۰هزار نفر جمعیت تقریباً ۱۳٪ مرگ‌های ناشی از سرطان گوارش و ۵/۳٪ از مرگ‌ومیر ناشی از علل غیر حوادث در ایران را شامل می‌شود (۶).

سرطان کولورکتال یکی از انواع سرطان‌هاست که شامل سرطان رودۀ بزرگ (سرطان کولون) و سرطان راست‌روده (سرطان رکتوم) و در اثر رشد کنترل نشدۀ لایه داخلی اندام‌های کولون و رکتوم ایجاد می‌گردد که با افزایش سن افراد، افزایش می‌یابد و این

سرطان با نام سرطان رودۀ بزرگ شناخته می‌شود (۸, ۷). بر اساس گزارش‌های، میزان بروز سرطان کولورکتال در ایران طی ۲۵ سال اخیر روند رو به رشدی داشته و در مقایسه با کشورهای غربی جمعیت جوان‌تری از کشور را تحت تأثیر خود قرار داده است (۹)، درنتیجه اهمیت بررسی آن به‌عنوان یک مسئله بهداشت عمومی حائز اهمیت است. بررسی مرگ‌ومیر بیماران مبتلا به سرطان، یکی از انواع مطالعاتی است که وضعیت بیماری و عوامل مرتبط با آن را مشخص می‌کند.

در سال‌های اخیر توجه فراوانی به مدل‌های آماری برای طبقه‌بندی داده‌های پزشکی با توجه به بیماری‌های مختلف و پیامدهای آنها شده است. طبقه‌بندی یکی از مهم‌ترین کاربرد روش‌های آماری در علوم مختلف است که هدف پیش‌بینی یک پاسخ دوحالتی یا چندحالتی بر اساس برخی متغیرهای مستقل در مورد موضوعات مختلف است (۱۰).

در آمار عمده روش‌های مدل‌سازی تعیین روابط بین متغیرها، تعیین متغیرهای اثرگذار و پیش‌بینی است. به دلیل پیش‌فرض‌های موجود برای مدل‌های آماری کلاسیک و پیچیدگی تفسیر نتایج آنها برای محققان پزشکی در دهه‌های اخیر تکنیک‌های دیگری رواج یافته است. جنگل تصادفی یکی از روش‌های آماری نوین محسوب می‌شود که محدودیت‌های مدل‌های آماری کلاسیک را ندارد.

جنگل تصادفی یک روش ناپارامتری و متعلق به خانواده‌ی روش‌های گروهی است که عملکرد امیدوارکننده و نویدبخشی را در مطالعات مختلف نشان می‌دهد (۱۱). از ویژگی‌های مهم جنگل تصادفی عملکرد بالای آن در اندازه‌گیری اهمیت متغیرها برای مشخص کردن اینکه هر متغیر چه نقشی در پیش‌بینی پاسخ دارد، است (۱۲).

بنابراین با توجه به مطالب مطرح‌شده و افزایش بروز سرطان کولورکتال در چند دهۀ گذشته (۱۳)، مطالعه حاضر باهدف پیش‌بینی مرگ‌ومیر حاصل از سرطان کولورکتال با استفاده از روش جنگل تصادفی انجام شد. همچنین ما عملکرد جنگل تصادفی را با رگرسیون لجستیک مقایسه کردیم.

روش کار

در این مطالعۀ گذشته‌نگر، پروندۀ ۳۰۴ بیمار مبتلا به سرطان کولورکتال که به بیمارستان طالقانی تهران مراجعه کرده بودند و اطلاعاتشان در پژوهشکده بیماری‌های گوارش و کبد دانشگاه علوم پزشکی شهید بهشتی ثبت‌شده بود، موردبررسی قرار گرفت. اطلاعات بیماران از سال ۱۳۸۸ تا ۱۳۹۳ ثبت‌شده بود. مرگ بیمار از طریق تماس با خانواده و اطرافیان بیمار موردبررسی قرار گرفت. این طرح در کمیته اخلاق دانشگاه علوم پزشکی شهید بهشتی تصویب‌شده بود. برای همه بیماران ویژگی‌های بالینی و جمعیت شناختی شامل سن در زمان تشخیص بیماری، جنسیت، سابقه خانوادگی سرطان، سن بیمار، محل تومور، مرحله سرطان، درجه تمایزیافتگی تومور، مکانیسم فرار از سلول ایمنی، سابقه بیماری دیابت، سابقه بیماری آنمی، سابقه بیماری IBD، نتایج MSI، جهش BRAF، جهش KRAS ، بتا کاتنین (B Catenin) و وضعیت حیاتی در تحلیل به کار رفت.

رگرسیون لجستیک[۱] یک ابزار تحلیلی بسیار عمومی است که اغلب در تحقیقات پزشکی برای پیش‌بینی متغیر پاسخ دوحالتی از طریق متغیرهای کمکی و عوامل، مورداستفاده قرار می‌گیرد (۱۴, ۱۵). برای یک متغیر پاسخ Y، یک بردار از متغیرها X و $\pi(X)$ که نشان‌دهنده شانس موفقیت با توجه به مقدار خاصی از X است، مدل رگرسیون لجستیک می‌تواند به‌صورت زیر نمایش داده شود.

$$logit[\pi(x)] = \log\left[\frac{\pi(x)}{1-\pi(x)}\right] = \alpha + \beta x$$

رگرسیون لجستیک از یک تبدیل لجیت برای محاسبه شانس یک متغیر پاسخ دوحالتی استفاده می‌کند (۱۵).

جنگل تصادفی[۲] یک روش یادگیری گروهی برای طبقه‌بندی است که توسط بریمن و همکاران در سال ۲۰۰۱ برای افزایش دقت طبقه‌بندی ارائه شد (۱۶). جنگل تصادفی یک طبقه‌بندی گروهی است که از تعدادی درخت تصمیم‌گیری تشکیل‌شده است که هر درخت با استفاده از یک نمونه بوت استرپ از داده‌های اصلی ساخته‌شده است. یک درخت به‌وسیله افراز بازگشتی نمونه بوت استرپ بر اساس بهینه‌سازی یک قانون تقسیم‌بندی، رشد می‌کند (۱۷). تجزیه‌وتحلیل جنگل تصادفی شامل ۱۰۰۰ درخت است که هرکدام شامل ۴ متغیر($\sqrt{P}$) (که P تعداد متغیرهای مستقل است) می‌باشند (۱۸). از روش k-fold برای بررسی اعتبار نتایج و مقایسه آن با افراد در رگرسیون لجستیک، استفاده شد. مجموعه داده آزمون شامل یک‌سوم (۳۳درصد) از کل داده‌ها و بقیه داده‌ها هم جزو مجموعه داده آموزشی می‌باشند. الگوریتم یادگیری را بر روی مجموعه داده آزمون اعمال می‌کنیم (به کار می‌بریم) و از مجموعه داده آموزشی برای الگوریتم تحت نظارت

[۱] Logistic Regression
[۲] Random Forest

استفاده می‌کنیم. تجزیه‌وتحلیل جنگل تصادفی با استفاده از بسته"randomForest" در نرم‌افزار R انجام شد. یک RF آن‌قدر بزرگ است که تفسیر آن کار بسیار دشواری است، لذا نیازمند خلاصه کردن اطلاعات آن با استفاده از شاخص‌های کمی هستیم. یکی از این شاخص‌ها اهمیت متغیر[۱] (VI) است. VI شاخصی برای رتبه‌بندی متغیرها برحسب اهمیت آنها در اثرگذاری روی پاسخ است. معروف‌ترین شاخص‌های VI، شاخص اهمیت جینی[۲] و شاخص اهمیت جایگشتی[۳] است (۱۹).

شاخص اهمیت جایگشتی: برای محاسبه این شاخص، الگوریتم RF از تمام مشاهدات نمونه برای ساخت درخت استفاده نمی‌کند بلکه یک نمونه تصادفی با جایگذاری به حجم n_1 (معمولاً برابر $\frac{2}{3}n$) از مشاهدات انتخاب می‌شود. به مشاهدات انتخاب‌شده نمونه آموزشی[۴] (LS) و به بقیه آنها نمونه خارج کیسه[۵] (OOB) گفته می‌شود. درخت‌ها با مشاهدات LS ساخته می‌شوند و از OOB برای اندازه‌گیری ناخالصی درخت استفاده می‌شود. در هر درخت ابتدا اندازه ناخالصی روی مشاهدات OOB محاسبه می‌شود. سپس مقادیر متغیر X_i مشاهدات OOB به‌طور تصادفی جابجا می‌شوند و اندازه ناخالصی درخت روی مقادیر جابجا شده محاسبه می‌شود. اندازه اهمیت متغیر X_i در هر درخت، اختلاف بین این دو اندازه ناخالصی است و میانگین این مقادیر شاخص اهمیت جایگشتی است. انگیزه این روش این است که اگر X_i متغیر مهمی باشد جابجا شدن مقادیر آن به‌طور تصادفی منجر به افزایش ناخالصی درخت می‌شود در حالی که اگر متغیر تأثیرگذاری نباشد، تغییری در ناخالصی ایجاد نمی‌شود (۱۹).

ازلحاظ تئوری(نظری)، جنگل تصادفی دارای چندین مزیت نسبت به رگرسیون است. مزیت اول، الگوریتم جنگل تصادفی می‌تواند متغیرهای مهم را به‌طور خودکار بدون توجه به اینکه چند متغیر در ابتدا استفاده‌شده، انتخاب کند که متفاوت از انتخاب مرحله‌ای(گام‌به‌گام) در رگرسیون لجستیک است. مزیت دوم، مقادیر گمشده و همچنین داده‌های نامتعادل می‌تواند به‌صورت خودکار توسط جنگل تصادفی به کار گرفته شود. مزیت سوم، جنگل تصادفی در مقایسه با رگرسیون لجستیک در مجموعه داده‌های بزرگ و با تعداد متغیر زیاد، عملکرد بهتری دارد (۲۰).

روش ارزیابی

برای مقایسه قدرت تمایز این دو مدل از مساحت زیر منحنی مشخصه عملکرد (AUC)، دقت (Accuracy)، ویژگی (Specificity) و حساسیت (Sensitivity) استفاده می‌کنیم. برای تجزیه‌وتحلیل داده‌ها از نرم‌افزار R استفاده شد. قبل از اینکه به این معیارها پرداخته شود لازم است به مفهوم ماتریس ابهام[۶] پرداخته شود. این ماتریس اجازه می‌دهد تجسمی از کارایی یک مدل ایجاد شود و حاوی اطلاعاتی در مورد دسته‌بندی واقعی و پیش‌بینی‌های انجام‌شده توسط یک سیستم دسته‌بندی است (۲۱). جدول شماره ۱ نشان‌دهنده ماتریس ابهام در یک مسئله دسته‌بندی دو کلاسه است.

جدول شماره ۱ – ماتریس ابهام

پیش بینی	داده‌های واقعی		مجموع کل
	مثبت	منفی	
مثبت	TP	FP	TP+FP
منفی	FN	TN	TN+FN
مجموع کل	TP+FN	FP+TN	n

در حوزه پزشکی منظور از مثبت، ابتلا به بیماری (در اینجا مرگ) و منظور از منفی، سالم (در اینجا زنده‌بودن) است.

- TP به معنی تعداد افرادی که مرده هستند و مدل پیش‌بینی کننده مرده پیش‌بینی کرده است.
- FP به معنی تعداد افرادی که زنده هستند و مدل پیش‌بینی کننده مرده پیش‌بینی کرده است.
- FN به معنی تعداد افرادی که مرده هستند و مدل پیش‌بینی کننده زنده پیش‌بینی کرده است.
- TN به معنی تعداد افرادی که زنده هستند و مدل پیش‌بینی کننده زنده پیش‌بینی کرده است.

با استفاده از مقادیر این ماتریس می‌توان مقدار معیارهای حساسیت[۷]، ویژگی[۸]، دقت[۹]، منفی کاذب[۱۰] و مثبت کاذب[۱۱] را

[۱] Variable Importance
[۲] Gini Importance Index
[۳] Permutation Importance Index
[۴] Learning Sample
[۵] Out-Of-Bag
[۶] Confusion matrix
[۷] Sensitivity
[۸] Specificity
[۹] Accuracy
[۱۰] False Negative Rate
[۱۱] False Positive Rate

محاسبه نمود. در ادامه به تعریف این معیارها می‌پردازیم.

حساسیت

منظور از معیار حساسیت نسبت تعداد افراد مرده‌ای است که به‌درستی مرده دسته‌بندی‌شده‌اند به‌کل افرادی که مرده‌اند. با توجه به اینکه هرچقدر مقدار این معیار بزرگ‌تر باشد نشان‌دهنده دقت تشخیص افراد مرده است، مقدار این معیار در حوزه پزشکی بسیار بااهمیت است. مقدار این معیار با استفاده از رابطه زیر قابل‌محاسبه است.

$$Sensitivity = \frac{TP}{FN + TP}$$

ویژگی

منظور از معیار ویژگی نسبت تعداد افراد زنده‌ای است که به‌درستی زنده دسته‌بندی‌شده‌اند به‌کل افراد زنده. این معیار با معیار حساسیت توازن[1] دارند. معیار ویژگی ازجمله معیارهایی است که در حوزه پزشکی دقت و کارایی مدل یادگیری را با استفاده از آن می‌سنجند و مقدار آن معیار با استفاده از رابطه زیر قابل‌محاسبه است.

$$Specificity = \frac{TN}{FP + TN}$$

دقت

منظور از معیار دقت نسبت تعداد افراد مرده و زنده‌ای است که به‌درستی دسته‌بندی‌شده‌اند به‌کل جمعیت. با توجه به اینکه دو معیار حساسیت و ویژگی با یکدیگر توازن دارند از این معیار به‌عنوان معیار تخمین استفاده می‌شود. مقدار این معیار با استفاده از رابطه زیر قابل‌محاسبه است.

$$Accuracy = \frac{TN + TP}{FP + FN + TN + TP}$$

منفی کاذب

منظور از معیار منفی کاذب نسبت تعداد افراد مرده‌ای است که زنده دسته‌بندی‌شده‌اند به‌کل افراد مرده. مقدار این معیار با استفاده از رابطه زیر قابل‌محاسبه است.

$$FNR = \frac{FN}{FN + TP}$$

مثبت کاذب

منظور از معیار مثبت کاذب نسبت تعداد افراد زنده‌ای است که مرده دسته‌بندی‌شده‌اند به‌کل افراد زنده. مقدار این معیار با استفاده از رابطه زیر قابل‌محاسبه است.

$$FPR = \frac{FP}{FP + TN}$$

برای تجزیه‌وتحلیل داده‌ها از نرم‌افزار R نسخه ۳ .۴ .۳ استفاده شد. تجزیه‌وتحلیل جنگل تصادفی با استفاده از بسته"randomForest" انجام شد. همچنین از بسته"ROCR" برای رسم منحنی مشخصه عملکرد و محاسبه شاخص‌های نیکویی برازش استفاده کردیم.

یافته‌ها

از تعداد ۳۰۴ بیمار مبتلا به سرطان کولورکتال ۱۵۶نفر (۵۱/۳ درصد) مرد و ۱۴۸نفر (۴۸/۷) زن بودند. میانگین سنی بیماران ۱۱/۱± ۵۶/۰۶ و حداقل و حداکثر سن به ترتیب ۲۸ و ۸۸ سال بود. در روش جنگل تصادفی بر اساس دو شاخص ضریب جینی و اهمیت جایگشتی، متغیرها بر پایه تأثیر آنها بر طبقه‌بندی مرتب شدند (۲۲). (شکل شماره ۱)

بر اساس شاخص ضریب جینی (سمت راست شکل ۱) سن در زمان تشخیص بیماری، سن بیمار، مرحله سرطان، درجه تمایزیافتگی تومور (Differentiation) و مکانیسم فرار از سلول ایمنی به ترتیب مهم‌ترین متغیرهای تأثیرگذار بر مرگ‌ومیر سرطان کولورکتال هستند. ترتیب مهم‌ترین متغیرهای تأثیرگذار بر اساس شاخص اهمیت جایگشتی (سمت چپ شکل ۱) هم به این صورت است: مرحله سرطان، سن بیمار، سن در زمان تشخیص بیماری، مکانیسم فرار از سلول ایمنی و درجه تمایزیافتگی تومور (Differentiation).

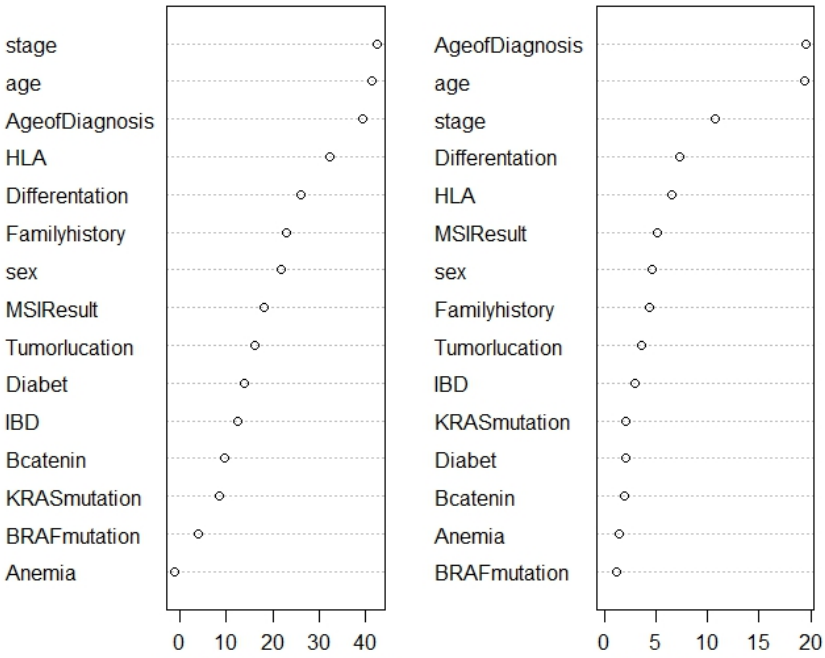
مدل رگرسیون لجستیک با استفاده از نرم‌افزار R شامل متغیرهای معنی‌دار حاصل از تحلیل تک متغیره، انجام شد. ورود متغیرها به مدل رگرسیون خطی تعمیم‌یافته بر اساس روش گام‌به‌گام(مرحله‌ای) و بهترین مدل با استفاده از معیار اطلاعات آکائیکی (AIC) انجام شد.

شکل شماره ۲ منحنی مشخصه عملکرد (ROC) برای روش‌های رگرسیون لجستیک و جنگل تصادفی را نشان می‌دهد. مساحت زیر منحنی مشخصه عملکرد برای روش‌های رگرسیون لجستیک و جنگل تصادفی به ترتیب برابر است با ۸۰٪ و ۹۸٪ که نشان‌دهنده

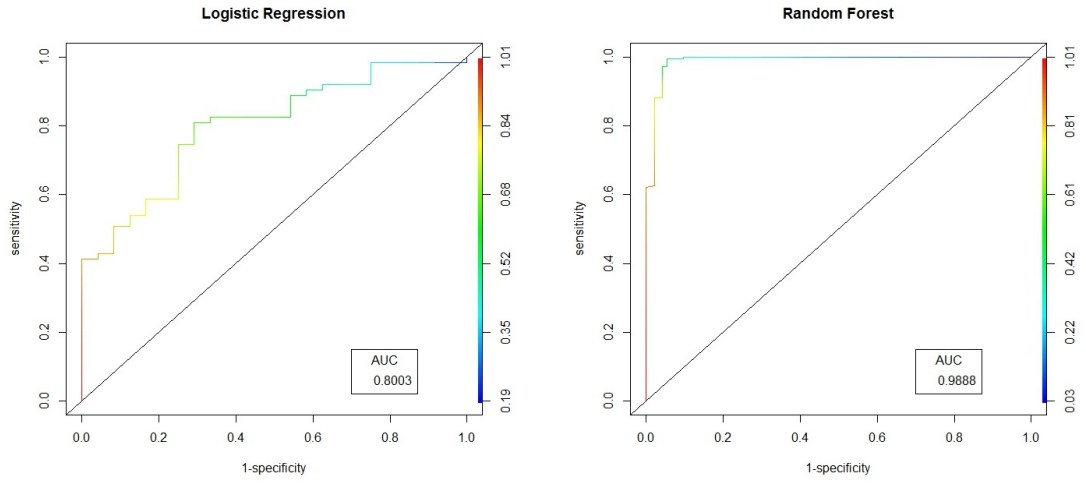
[1] Trade off

عملکرد بهتر روش جنگل تصادفی است

اطلاعات در مورد ویژگی، حساسیت، دقت، درصد منفی کاذب و درصد مثبت کاذب برای روش‌های انجام‌شده در جدول شماره ۲ نشان داده‌شده است که عملکرد بهتری برای روش جنگل تصادفی در مقایسه با روش رگرسیون لجستیک به دست آمد.



شکل شماره ۱ – شاخص ضریب جینی و شاخص اهمیت جایگشتی



شکل شماره ۲ – منحنی‌های مشخصه عملکرد برای روش‌های جنگل تصادفی و رگرسیون لجستیک

جدول شماره ۲ – دقت، ویژگی و حساسیت برای روش‌های رگرسیون لجستیک و جنگل تصادفی

شاخص‌های نیکویی برازش	جنگل تصادفی	رگرسیون لجستیک
حساسیت (Sensitivity)	٪۸۳	۶۰٪
ویژگی (Specificity)	۸۶٪	۷۱٪
دقت (Accuracy)	۸۵٪	۶۲٪
منفی کاذب (False Negative Rate)	۱۷٪	۴۰٪
مثبت کاذب (False Positive Rate)	۱۴٪	۲۹٪

بحث

افزایش سرعت پیر شدن جوامع در کشورهای صنعتی، تعداد موارد سرطان ازجمله سرطان کولورکتال را به‌سرعت افزایش می‌دهد. بروز سرطان کولورکتال در سه دههٔ اخیر روند صعودی را نیز در ایران نشان می‌دهد (۲۳). سرطان کولورکتال سالانه با تعداد نزدیک به یک‌میلیون مورد جدید و مرگ حدود ۵۰ درصد مبتلایان در پنج سال اول شروع بیماری به‌عنوان مشکل مهمی در سلامت عمومی شناخته می‌شود (۲۴). همین امر ضرورت مطالعه در مورد این سرطان را ایجاد می‌کند. در این مطالعه، بعضی از عوامل که ممکن است در مرگ‌ومیر بیماران سرطان کولورکتال مؤثر باشند ازجمله سن در زمان تشخیص بیماری، جنسیت، سابقه خانوادگی سرطان، سن بیمار، محل تومور، مرحله سرطان، درجه تمایزیافتگی تومور، مکانیسم فرار از سلول ایمنی، سابقه بیماری آنمی، سابقه بیماری IBD، نتایج MSI، جهش BRAF، جهش KRAS، بتا کاتنین (B Catenin)، سابقه بیماری دیابت و وضعیت حیاتی مورد ارزیابی قرار گرفتند. از دو روش جنگل تصادفی و رگرسیون لجستیک استفاده کردیم. نتایج مطالعه ما نشان داد که ازنظر معیارهای مختلف روش جنگل تصادفی دارای عملکرد بهتری نسبت به روش رگرسیون لجستیک بود. در مطالعه نوری و همکاران (۲۵) که از روش آماری ناپارامتری و نوین جنگل تصادفی استفاده کردند، بیان کردند که در تعیین متغیرهای مهمی که روی پاسخ تأثیرگذارند، روش جنگل تصادفی استنباط بهتری از متغیرها را می‌دهد.

با توجه به شکل شماره ۱، تجزیه‌وتحلیل اطلاعات نشان داد که متغیر سن در زمان تشخیص بیماری تأثیر بسزایی بر مرگ‌ومیر افراد مبتلا به سرطان کولورکتال داشت. مطالعه نصیری و همکاران (۲۶) رابطه معنی‌داری را بین سن در زمان تشخیص و بقای بیماران تأیید کرده‌اند به‌طوری‌که میانگین بقای افراد در گروه سنی زیر ۶۵ سال به‌طور معنی‌داری بیشتر از میانگین بقای افراد در گروه سنی بالای ۶۵ سال را گزارش داده است. همچنین در مطالعه ساکی مالحی و همکاران (۲۷) تأثیر سن در زمان تشخیص بر بقای بیماران تأیید شد. در مطالعاتی که توسط لیانگ و همکاران (۲۸) و چین و همکاران (۲۹) انجام‌شده است، سن در زمان تشخیص به‌عنوان عاملی مهم در مرگ بر اثر سرطان کولون گزارش‌شده است.

همچنین مطالعه‌ی ما نشان داد که مرحله بیماری بر خطر مرگ افراد در اثر سرطان کولورکتال، تأثیر معنی‌داری داشت. در مطالعه‌ی نصیری و همکاران (۲۶) مرحله بیماری عامل مهمی در پیش‌آگهی بیماران بود به‌طوری‌که بقا بیماران با بالا رفتن مرحله بیماری کاهش می‌یابد و هر چه مرحله‌ی بیماری پایین‌تر باشد، بقا بیماران بهتر خواهد بود. همچنین در مطالعه دهکردی و همکاران (۳۰) تأثیر مرحله‌ی بیماری بر بقا بیماران تأیید شد.

در مطالعه ما درجه تمایزیافتگی تومور اثر معنی‌داری بر مرگ افراد در اثر سرطان کولورکتال داشت. در مطالعه‌ای که توسط لیانگ و همکاران (۲۸) انجام‌شده است ارتباط معنی‌داری میان بقا بیماران مبتلا به سرطان کولون و درجه تمایزیافتگی مشاهده گردید. همچنین در مطالعات روشنایی و همکاران (۳۱)، آخوند و همکاران (۳۲) و دهکردی و همکاران (۳۳) تأثیر درجه تمایزیافتگی تومور بر بقا معنی‌دار بود.

از نقاط قابل‌توجه این مطالعه مقایسه دو روش آماری جنگل تصادفی و رگرسیون لجستیک است که می‌توان به توانایی بالاتر روش جنگل تصادفی نسبت به روش پرکاربرد رگرسیون لجستیک اشاره کرد و از محدودیت‌های این مطالعه می‌توان به عدم دسترسی به برخی اطلاعات بیماران اشاره کرد.

نتیجه‌گیری

در این مطالعه از دو روش جنگل تصادفی و رگرسیون لجستیک برای پیش‌بینی مرگ‌ومیر بیماران مبتلا به سرطان کولورکتال و تعیین عوامل خطر مرتبط با آن استفاده شد. بر اساس معیارهای حساسیت، ویژگی، دقت و مساحت زیر منحنی ROC روش جنگل تصادفی از توانایی بالاتری نسبت به رگرسیون لجستیک برخوردار بود؛ بنابراین روش جنگل تصادفی ابزار مناسب‌تری برای پیش‌بینی مرگ‌ومیر بیماران مبتلا به سرطان کولورکتال در این گونه داده‌ها است. باوجوداینکه مطالعات متعددی در حوزه‌ی سرطان کولورکتال در سراسر جهان انجام‌شده، ولی صحت عوامل تعیین‌شده هنوز هم جای بحث دارد. چون با توجه به مطالعات دیگر، برخی از عوامل مؤثر بر مرگ‌ومیر بیماران مبتلا به سرطان کولورکتال به‌صورت مختلف و گوناگون گزارش‌شده است، لذا انجام مطالعات بیشتر جهت تعیین عوامل مؤثر بالینی و جمعیت شناختی بر مرگ‌ومیر بیماران مبتلا به سرطان کولورکتال می‌تواند مفید باشد. چون سن در زمان تشخیص و مرحله بیماری به‌عنوان عامل مؤثر در مرگ‌ومیر معرفی شد و مراحل بالای این بیماری می‌تواند با افزایش خطر مرگ همراه باشد، بنابراین آگاهی دادن به جامعه در جهت مراجعه هر چه سریع‌تر به پزشک و انجام معاینات، لازم و ضروری به نظر می‌رسد. همچنین با تشخیص زودرس سرطان با برنامه‌های بیماریابی و غربالگری مفید می‌توان بر طول عمر بیماران افزود.

منابع

1. Azeem S, Gillani SW, Siddiqui A, Jandrajupalli SB, Poh V, Syed Sulaiman S. Diet and colorectal cancer risk in Asia-A systematic review. Asian Pac J Cancer Prev. 2015; 16: 5389-96.
2. Jemal A, Siegel R, Ward E, Hao Y, Xu J, Murray T, et al. Cancer statistics, 2008. CA: a cancer journal for clinicians. 2008; 58: 71-96.
3. Ahmadi A, Mosavi-Jarrahi A, Pourhoseingholi MA. Mortality determinants in colorectal cancer patients at different grades: a prospective, cohort study in Iran. Asian Pacific journal of cancer prevention: APJCP. 2015; 16: 1069-72.
4. Rezaianzadeh A, Safarpour AR, Marzban M, Mohaghegh A. A systematic review over the incidence of colorectal cancer in Iran. Annals of colorectal research. 2015; 3.
5. Sadjadi A, Malekzadeh R, Derakhshan MH, Sepehr A, Nouraie M, Sotoudeh M, et al. Cancer occurrence in Ardabil: Results of a population-based Cancer Registry from Iran. International journal of cancer. 2003; 107: 113-8.
6. Ganji A, Safavi M, Nouraie S, Nasseri-Moghadam S, Merat S, Vahedi H, et al. Digestive and liver diseases statistics in several referral centers in Tehran, 2000-2004. Govaresh. 2006; 11: 3-8.
7. Baghestani AR, Daneshvar T, Pourhoseingholi MA, Asadzade H. Survival of colorectal cancer patients in the presence of competing-risk. Asian Pac J Cancer Prev. 2014; 15: 6253-5.
8. Li C, Lu H-J, Na F-F, Deng L, Xue J-X, Wang J-W, et al. Prognostic role of hypoxic inducible factor expression in non-small cell lung cancer: a meta-analysis. Asian Pacific Journal of Cancer Prevention. 2013; 14: 3607-12.
9. Bishehsari F, Mahdavinia M, Vacca M, Malekzadeh R, Mariani-Costantini R. Epidemiological transition of colorectal cancer in developing countries: environmental factors, molecular pathways, and opportunities for prevention. World journal of gastroenterology: WJG. 2014; 20: 6055.
10. Duda RO, Hart PE, Stork DG. Pattern classification: John Wiley & Sons; 2012.
11. Breiman L. Random forests. Machine learning. 2001; 45: 5-32.
12. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. The annals of applied statistics. 2008: 841-60.
13. Ansari R, Amjadi H, Norozbeigi N, Zamani F, Mir-Nasseri S, Khaleghnejad A, et al. Survival analysis of colorectal cancer in patients underwent surgical operation in Shariati and Mehr Hospital-Tehran, in a retrospective study. Govaresh. 2007; 12: 7-15.
14. Harrell FE. Ordinal logistic regression. Regression modeling strategies: Springer; 2001, 331-43.
15. Hosmer Jr DW, Lemeshow S, Sturdivant RX. Applied logistic regression: John Wiley & Sons; 2013.
16. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: a classification and regression tool for compound classification and QSAR modeling. Journal of chemical information and computer sciences. 2003; 43: 1947-58.
17. Tufféry S. Data mining and statistics for decision making: Wiley Chichester; 2011.
18. Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. BMC bioinformatics. 2008; 9: 307.
19. Breiman L. Classification and regression trees: Routledge; 2017.
20. Geng M. A comparison of logistic regression to random forests for exploring differences in risk factors associated with stage atdiagnosis between black and white colon cancer patients: University of Pittsburgh; 2006.
21. Kohavi R. Glossary of terms. Special issue on applications of machine learning and the knowledge discovery process. 1998; 30: 127-32.
22. Boulesteix A-L, Bender A, Lorenzo Bermejo J, Strobl C. Random forest Gini importance favours SNPs with large minor allele frequency: impact, sources and recommendations. Briefings in Bioinformatics. 2011; 13: 292-304.
23. Hosseini SV, Izadpanah A, Yarmohammadi H. Epidemiological changes in colorectal cancer in Shiraz, Iran: 1980– 2000. ANZ journal of surgery. 2004; 74: 547-9.
24. 24. Newton K, Newman W, Hill J. Review of biomarkers in colorectal cancer. Colorectal disease. 2012; 14: 3-17.
25. Noori s, Nourijelyani k, Mohammad k, Niknam M, Mahmoudi M, Andonian L, et al. Random forests analysis: A modern statistical method for screening in high-dimensional studies and its application in a population-based genetic association study %J Journal of North Khorasan University of Medical Sciences. 2012; 3: 93-101.
26. Nasiri S, Sorush A, Karamnezhad M, Mehrkhani F, Mosafa S, Hedayat A. Prognostic Factors in the Survival Rate of Colorectal Cancer Patients after Surgery. 2010.

27. Saki Malehi A, Hajizadeh E, Fatemi R. Evaluation of prognostic variables for classifying the survival in colorectal patients using the decision tree. Iranian Journal of Epidemiology. 2012; 8: 13-9.
28. Liang H, Wang XN, Wang BG, Pan Y, Liu N, Wang DC, et al. Prognostic factors of young patients with colon cancer after surgery. World Journal of Gastroenterology: WJG. 2006; 12: 1458.
29. Chin CC, Wang JY, Yeh CY, Kuo YH, Huang WS, Yeh CH. Metastatic lymph node ratio is a more precise predictor of prognosis than number of lymph node metastases in stage III colon cancer. International journal of colorectal disease. 2009; 24: 1297-302.
30. MOGHIMI DB, SAFAEI A, ZALI MR. Survival rates and prognostic factors in colorectal cancer patients. 2008.
31. Roshanaei G, Komijani A, Sadighi A, Faradmal J. Prediction of survival in patients with colorectal cancer referred to the Hamadan MRI center using of Weibull parameter model and determination of its risk factors during 2005-2013. 2014.
32. Akhoond M, Kazemnejad A, Hajizadeh E, Fatemi S, Motlagh A. Investigation of Influential Factors Affecting Survival Rate of Patients with Colorectal Cancer using Copula Function. Iranian Journal of Epidemiology. 2011; 6: 40-9.
33. Moghimi-Dehkordi B, Safaee A, Zali MR. Prognostic factors in 1,138 Iranian colorectal cancer patients. International journal of colorectal disease. 2008; 23: 683-8.

Iranian Journal of Epidemiology Winter 2019; 14 (4): 375-383.
Original Article

Comparison of Random Forest and Logistic Regression Methods in Predicting Mortality in Colorectal Cancer Patients and Its Related Factors

Keshtvarz Hesam Abadi AM[1], Hajizadeh E[2], Pourhoseingholi MA[3], Nazemalhossein Mojarad E[4]

1- Masters Student of Biostatistics, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran
2- Professor of of Biostatistics, Faculty of Medical Sciences, Tarbiat Modares University,Tehran, Iran
3- Gastroenterology and Liver Diseases Research Center, Research Institute for Gastroenterology and Liver Diseases, Shahid Beheshti University of Medical Sciences, Tehran, Iran
4- Gastroenterology and Liver Disease Research Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran
Corresponding author: Hajizadeh E, hajizade@modares.ac.ir
(Received 1 June 2018; **Accepted** 29 September 2018)

Background and Objectives: The purpose of this study was to predict the mortality rate of colorectal cancer in Iranian patients and determine the effective factors on the mortality of patients with colorectal cancer using random forest and logistic regression methods.

Methods: Data from 304 patients with colorectal cancer registry from the Gastroenterology and Liver Research Center of Shahid Beheshti University of Medical Sciences during the years 2009 to 2014 were used as a retrospective study. Data analysis was performed using random forest and logistic regression methods. To analyze the data, R software version 3.4.3 was considered.

Results: Ten important variables related to colorectal cancer deaths were selected by random forest method. Several criteria such as the area under the characteristic curve (AUC) were used to compare the random forest method with logistic regression. According to both criteria, five important variables ranked by random forest were Cancer stage, age of diagnosis, patient's age, HLA, and degree of differentiation (tumor differentiation). In terms of different criteria, the random forest method had better performance than logistic regression (Area under the ROC curve for random forest and logistic regression methods was: 98%; 80% respectively).

Conclusion: Variables such as Cancer stage, age of diagnosis, patient's age, HLA, and degree of differentiation are considered as the most important factors affecting mortality in colorectal cancer, that the patients' longevity can be increased with the early diagnosis of cancer and screening programs.

Keywords: Colorectal cancer, Random forest, Logistic regression, Mortality